論　　文

データマイニングシステムの可用性テスト

Performance Tests through One Year Operation
of a Data Mining System

高　橋　道　郎
Michio TAKAHASHI
永　井　義　明
Yoshiaki NAGAI

１．はじめに

ウェブ上の記事や論説を計数・分析するデータマイニングシステムを開発し運用中である。このシステムは技術革新や政策等の外部環境の変化がベンチャー企業の消長にどのように影響を及ぼしているかを研究するに当たり，外部環境の測定手段として開発したものである。昨年は「SARS」「ベンチャー」「リストラ」「起業」とのキーワードを採りあげ計測した[1)]が，同一条件にて更に１年の計測を継続したのでその結果を報告する。

２．データマイニング装置の構成と機能

2.1　システム概要[2) 3) 4)]

あらかじめ登録したＷＥＢサイト（20箇所程度）を巡回して記事を検索し，この記事をもとにテキストマイニングを行うことにより，キーワード，論説等が時間の経緯とともにどう変化するのか，ビジネスとの関わりはどうかなどを調査検討するための研究プラットフォームである。詳細については報告済みであるので概要のみ以下に述べる。

このシステムの持つ機能は次の通りである。

- 指定されたＷＥＢサイトを巡回収集するロボット機能
- キーワードの表記上の振れを検索可能なあいまい検索機能
- 検索にヒットした記事リストの出力，記事数の推移のグラフ出力を行うレポート機能

2.2 機能構成

本システムの機能構成を図１に示す。

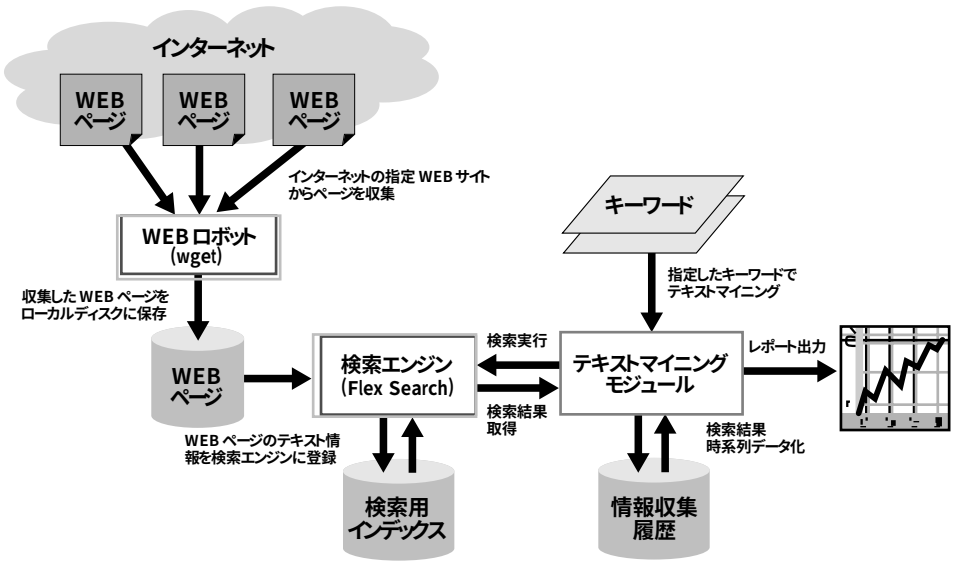


図１　ITプラットフォームシステム（Phase1）の機能構成

ここで,「ＷＥＢロボット」は

(1) ＷＥＢページ収集,
(2) ＷＥＢページ収集ツール,
(3) ＷＥＢ同期

とから成り,
「検索エンジン」は

(1) キーワード検索,
(2) キーワード検索エンジン

とから成り,
「テキストマイニングモジュール」は

(1) 記事抽出,
(2) 記事収集履歴管理,
(3) レポート出力

とから成っている。

３．運　用

2004年１月より2005年１月までの13ヶ月間にわたり，キーワードとして「SARS，リストラ，ベンチャー，起業，技術，中国」の６つを選定し，そのキーワードの現れる記事数を計数した。このうち「SARS，リストラ，ベンチャー，起業」の４つは昨年の研究と同じ語とし，比較検討ができるようにした。対象としたサイトは以下に示すものである。

表１．収集対象としたサイトの一覧

対象サイト名	ＵＲＬ
日本経済新聞社	http://www.nikkei.co.jp/
朝日新聞社	http://www.asahi.com/
讀売新聞社	http://www.yomiuri.co.jp/
共同通信社	http://www.kyodo.co.jp/
時事通信社	http://www.jiji.co.jp/
ロイター通信社	http://www.reuters.co.jp/
ニューズウィーク日本	http://www.nwj.ne.jp/
日経産業新聞	http://ss.nikkei.co.jp/ss/
ニフティ	http://www.nifty.com/category/news/
ヤフー	http://www.yahoo.co.jp/hl
産経新聞社	http://www.sankei.co.jp/
東洋経済新報社	http://www.toyokeizai.co.jp/

４．実験結果

以下はこれらのサイトにキーワード「SARS，リストラ，ベンチャー，起業，技術，中国」を含む記事が合計何個あるかを日毎に計測した値である。横軸は日，縦軸はこれらのキーワードを含む記事の数である。「SARS，リストラ，ベンチャー，起業」については前回の研究で興味深い結果が得られたので調査を継続することとしたものである。また「技術，中国」の２つの語についてはより一般的な語を選定した場合についての知見を得ることが目的である。

まず始めに新型肺炎ＳＡＲＳをキーワードとした計測結果を図２，図３に示す。

図２　SARSをキーワードとした計測結果（2003年５月～11月）

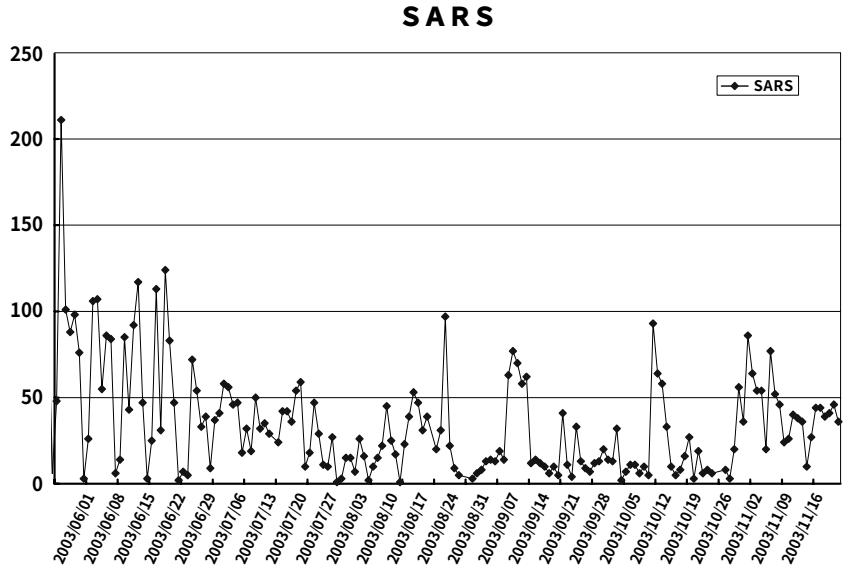


図３　SARSをキーワードとした計測結果（2004年１月～2005年１月）

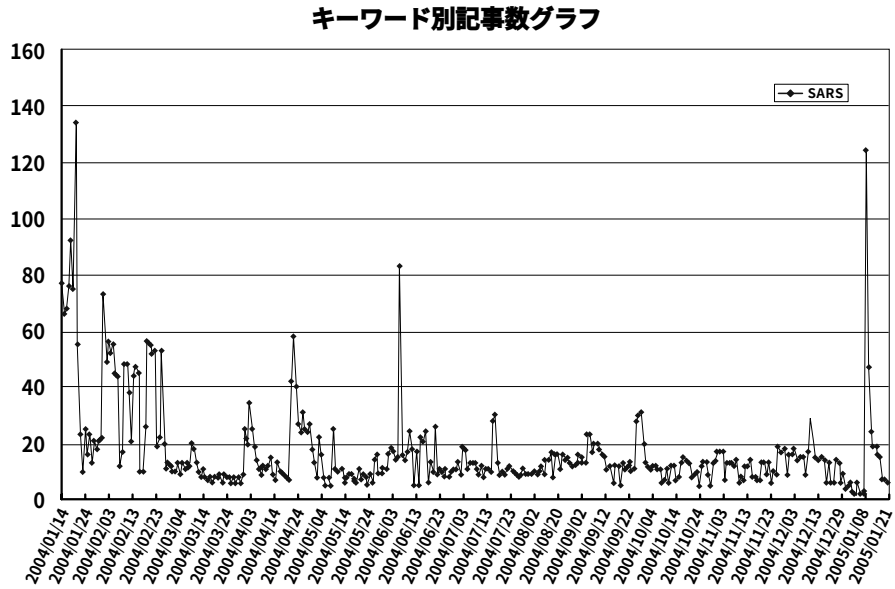


図２に関しては前報[1)]に次の解説を付けた。少し長くなるが引用すると“ SARSは2003年正月から春先にかけて猛威をふるっていたが夏には沈静化した。記事数をみると６月と７，８月の間には明らかな差があり，記事数も沈静化していることがよく

分かる。11月になり寒気をむかえると記事が増えているのも興味深い。また，６月に顕著であるが曜日によって大きな差があるのは，曜日と記事の編集方針に大きな相関があることを表していると考えられる。ちなみに６月の記事数の谷はいずれも週末である。”

図２と同じ測定を2004年１月～2005年１月まで継続したものを図３に示した。冬季には多かった記事数が３月にはいると減少，６月から12月までは10ないし20ヒットを保ち，12月に入り再び上昇に転じている。SARSに関する社会的な関心の強弱をそのまま反映しているヒット数といえる。同時にSARSという用語がこの２年間，同じ意味で使われ代替用語が出現してないことも伺われる。

次に先回の調査で採りあげた「ベンチャー」について述べる。

図４に，2003年５月～11月までのキーワード「ベンチャー」を持つ記事の増減を，図５に2004年１月～2005年１月までの分を掲げた。図４の2003年度データは期間中に２つの山を持ちその説明がつかなかったが，図５の2004年度に於いても不可解な点がある。それは３月から４月にかけてヒット数が不連続的に半減していることである。このデータマイニングシステムは対象となるウェブページを調べ，該当する用語やその類義語を持つ記事を計数している。仮に新聞などが連載企画記事を掲載し，その主題がキーワードに関係が深いとすれば，年度変わりによる連載終了に伴ってこのような急減が発生することは不思議ではない。しかし，この解釈が事実とすれば，このデータマイニングシステムの出力が社会事象を反映するとの仮説が揺らぐこととなり，能力の限界を露呈したとも言えよう。

図４　キーワード「ベンチャー」を持つ記事の増減（2003年５月～11月）

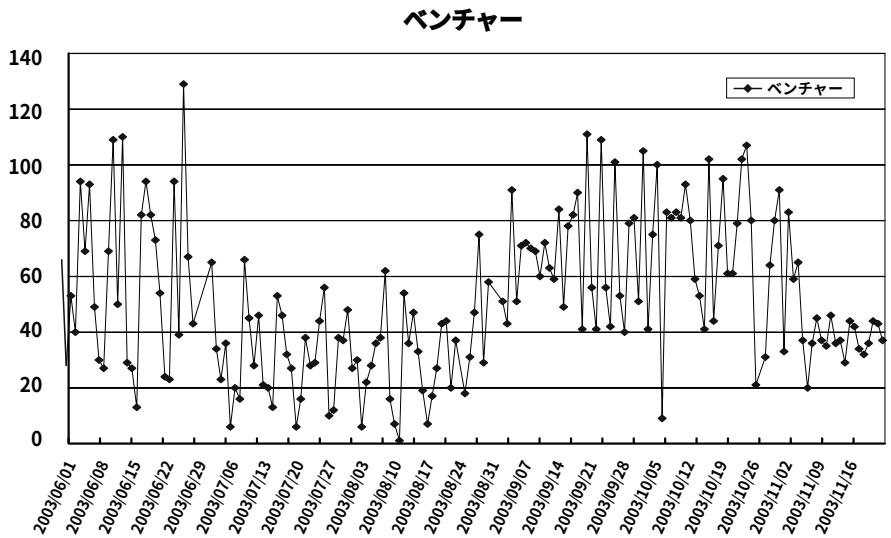
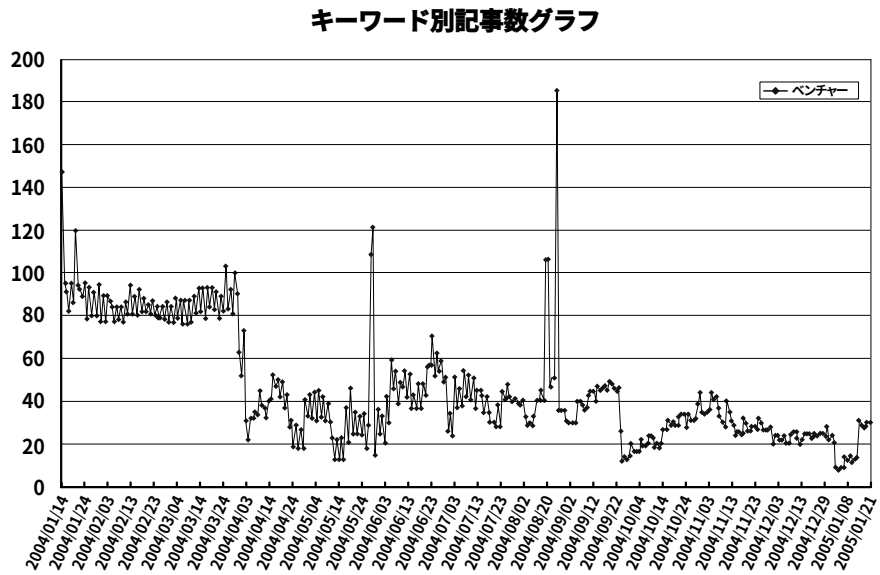


図５　キーワード「ベンチャー」を持つ記事の増減（2004年１月～2005年１月）



さて図６，図７にはよりヒット数の少ない「リストラ」をキーワードにした計測結果を示した。

2003年度データの解釈では“７月と８月には少なく，９月末に著しく多い日が続く”と述べたが図７に示した2004年度データに於いても９月に大きな山がある。だとすれば，年中行事としてこの時期にはこの話題で賑わうと解釈せざるをえない。経済事象を理解するために各種統計から季節変動を除いて判断をしなければならないように，キーワードヒット数の分析においても同様の配慮が求められることを示唆している。

図６　「リストラ」をキーワードにする記事の増減（2003年５月～11月）

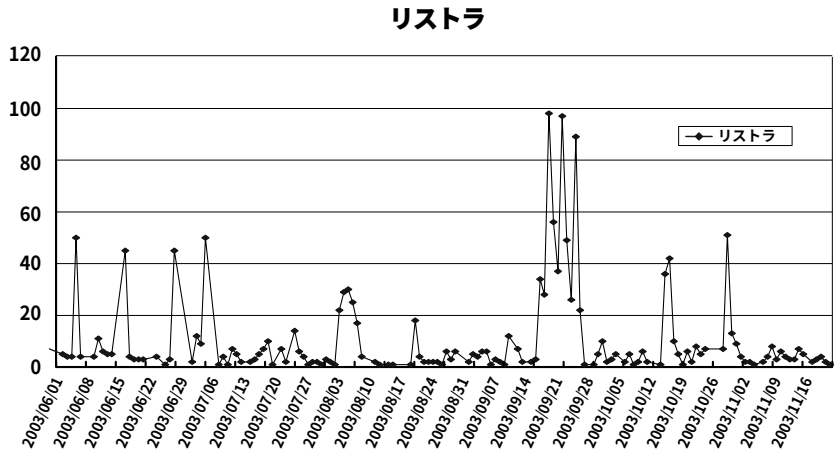
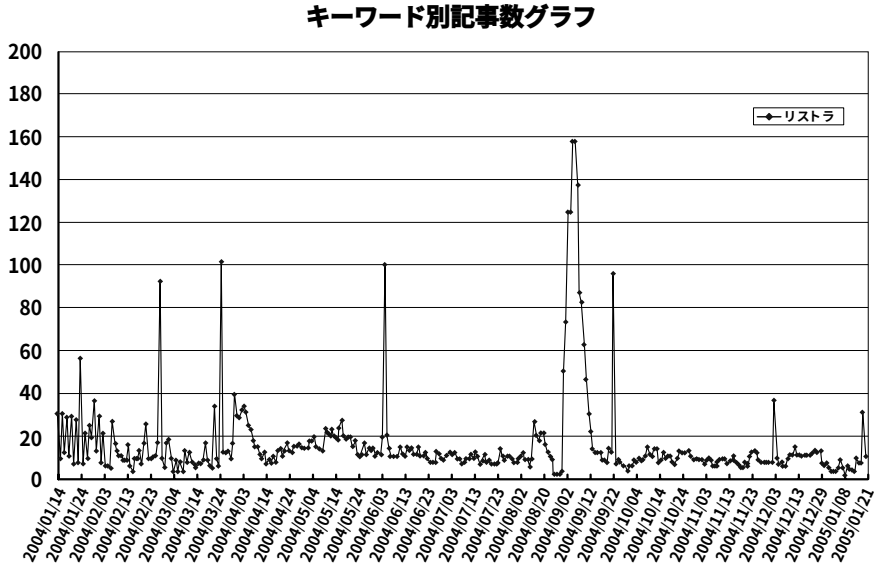


図７「リストラ」をキーワードにする記事の増減（2004年１月～2005年１月）



さて「リストラ」と同じようにヒット数の少ない「起業」について見ていこう。

図８が2003**年５月～**11**月の「起業」ヒット数，図９が**2004**年１月～**2005**年１月のヒット数である。**

図８　「起業」をキーワードにする記事の増減（2003年５月～11月）

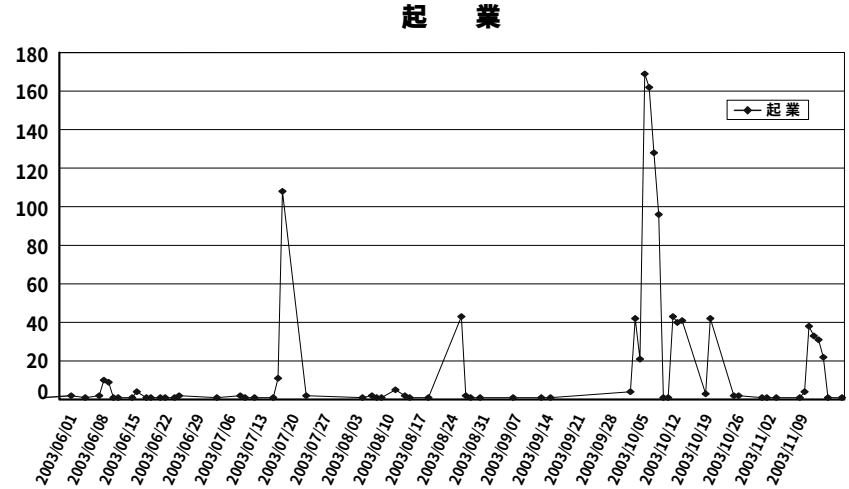
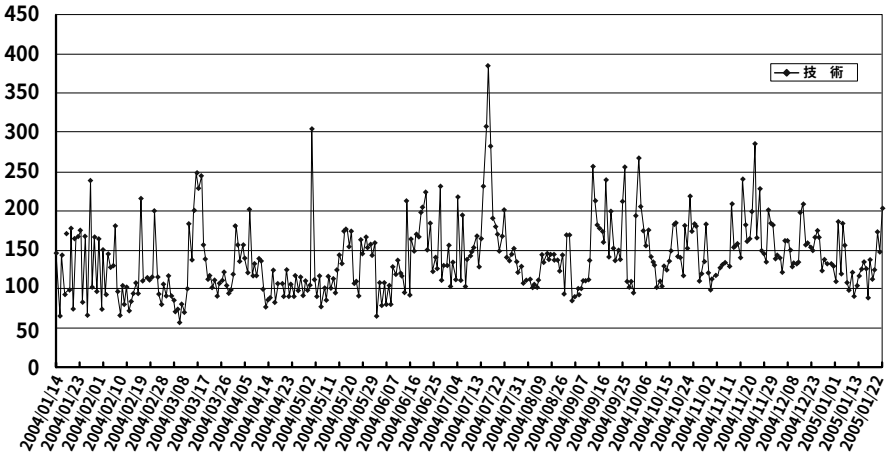


図９「起業」をキーワードにする記事の増減（2004年１月～2005年１月）

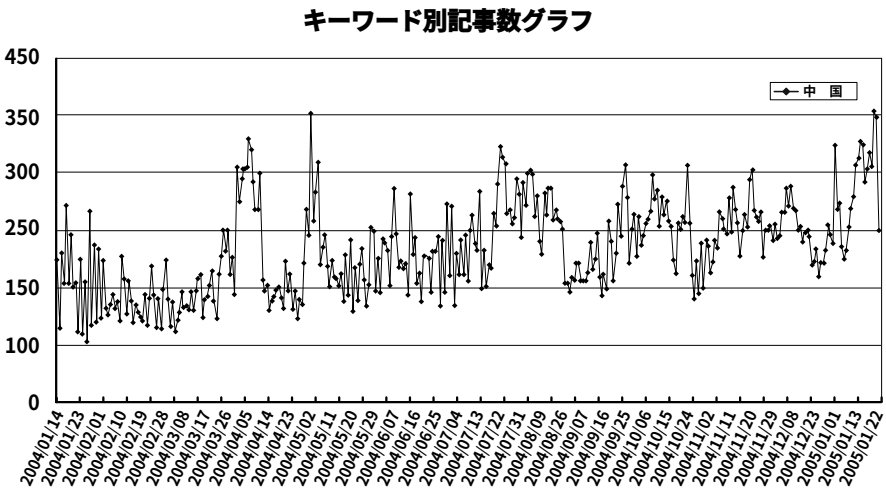
2004年３月まではヒット数のベースラインは０であったものが，４月以降は10程度に上昇している。これは図５の解釈で述べた「連載記事終了」の裏返しである「連載記事やキャンペーンの開始」との解釈が可能である。もっと素直に，起業が盛んになった，起業に関しての社会的な関心が増した，と主張もできる。今後の課題である。

図10 「技術」をキーワードにする記事の増減（2004年１月～2005年１月）



これまで掲げた４つの例は平均的ヒット数が100に満たないものばかりであった。このような例では，新聞の連載記事開始などの人為的な事柄に計測結果が大きく影響され，社会事象の計測としては信頼度にかけるとの批判が出るのもやむを得ない。そこで使用頻度の大きいであろう用語を選び計測結果を分析する。用いたキーワードは「技術」と「中国」である。図10は「技術」をキーワードにする記事の増減（2004年１月～2005年１月），図11は「中国」をキーワードにする記事の増減（2004年１月～2005年１月）である。

図11 「中国」をキーワードにする記事の増減（2004年１月～2005年１月）



「技術」ヒット数は100ないし200程度でほぼ横ばいである。「中国」ヒット数は200から300件であり少しずつ増加していると言える。中国や技術に関しては新聞等での特集記事，連載記事も多々存在するであろうが，それらが平均化されてヒット数としては顕著な変化としては現れていないと考えられる。社会事象をデータマイニングといったキーワード検索から推定しようとのアプローチを採るならば新聞社などメディアが行う個々の企画が平均化されるように調査を設計することが重要である。その目安はヒット数100と言える。

５．考　察

SARSの項で述べたように計測結果は社会的な関心の強弱をそのまま反映していると言える。しかし，ヒット数が小さいものにおいては新聞などの連載企画記事に計測結果が大きく影響されているらしいことも明らかになった。このことはデータマイニングシステムの出力が社会事象を反映するとの仮説が揺らぐこととなり，能力の限界を露呈したとも言えよう。

また，「リストラ」の項で述べたように年中行事として特定の時期に特定の話題がもてはやされる現象があることも浮き彫りになった。経済事象を理解するために各種統計から季節変動を除いて判断しなければならないように，キーワードヒット数の分析においても同様の配慮が求められることを示唆している。

２年間の調査をした４つの例は平均的ヒット数が100に満たないものばかりであった。このような例では，新聞の連載記事開始などの人為的な事柄に計測結果が大きく影響され，社会事象の計測としては信頼度にかけるとの批判が出るのもやむを得ない。社会事象をデータマイニングといったキーワード検索から推定しようとのアプローチを採るならば新聞社などメディアが行う個々の企画が平均化されるように調査を設計する上で，ヒット数100は一つの目安になる値と言える。

昨年も述べたように，用語と対象サイトを適切に選定することと，数年にわたる継続的な変化を追跡することが重要と考えられる。

６．結　論

開発したデータマイニングシステムを用いて調査を計測した。2004年１月から2005年１月までの13ヶ月の運用を行い

（１）　安定稼働に問題はない。
（２）　計測結果は社会的な関心の強弱をそのまま反映していると言える。
（３）　ヒット数が小さいものにおいては新聞などの連載企画記事に計測結果が大きく影響されているらしいことも明らかになった。
（４）　特定の時期に特定の話題が特に多く採りあげられる現象を確認した。経済事象を理解するために各種統計から季節変動を除いて判断しなければならないように，キーワードヒット数の分析においても同様の配慮が求められる。
（５）　調査を設計する上で，ヒット数100は一つの目安である。

との結論に至った。

また，用語と対象サイトを適切に選定することと，数年にわたる継続的な変化を追跡することが重要と考えられる。

参考文献

1） 高橋・永井，技術革新や政策等の外部環境とベンチャー企業の消長の相関－データマイニングシステムの構想と運用試験－，中部大学産業経済研究所紀要　第14号　pp.85－93。

2）「ＩＴプラットフォームP1基本設計書」，2003.3。

3）「ＩＴプラットフォームP1詳細設計書」，2003.3。

4）「ＩＴプラットフォームP1取扱説明書」，2003.3。